BayesCube v3.0 manual

Contents

1 Bayesian network editor

1.1 Creating a new BN model

To start the construction of a new Bayesian network model, select the FILE—NEW—MODEL XML menu item or the Model XML icon in the menu bar.

1.2 Opening a BN model

To open an existing BN model, select the FILE—OPEN menu item or the File open icon in the menu bar, and select the path in the dialogue box.

1.3 Definition of variable types

Variable types support the construction of similar variables. From a technical point of view, the use of variable types allow the definition of types for different arity, i.e. for binary, tertiary, quaternary variables (2,3,4-valued variables). However, variable types can be also used to express common semantic properties of variables, e.g. the values of propositional variables can be named TRUE/FALSE.

The variable type defines the nominal values of a discrete variable, its dialogue box can be opened by right clicking in the main window (in Editor mode) and selecting the item VARIABLE TYPES... from the pop-up menu. In the dialogue box new type can created and existing types can be modified. To create a new type, click on ADD NEW TYPE, rename it, and click on the ADD NEW VALUE to create the necessary number of values. The name of the nominal value, potential real values indicating lower and upper bounds can be specified in the list of values (these parameters do not influence the inference). Free text annotations with keywords can be created for the selected type by clicking ANNOTATION button.

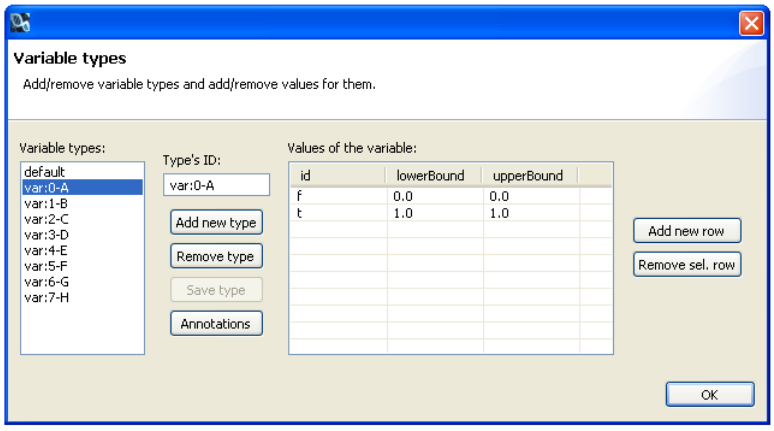

Figure 1: The dialogue box for defining variable types.

The DEFAULT type is not modifiable. The usage of types is described in Section 1.6.

1.4 Definition of variable groups

Variable groups support the visualization of the correspondence of variables with common semantic properties, e.g. the same visual appearance can be defined for the nodes representing random variables with common functional properties. The VARIABLE GROUPS dialogue box can be opened by right-clicking in the main window and selecting the VARIABLE GROUPS ... item. For a group, (1) name, (2) color and (3) its annotation list can be edited, which contains keyword-free text pairs.

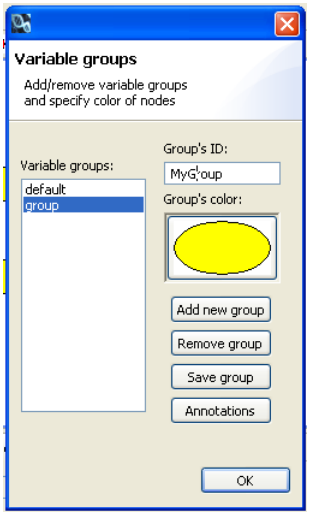

Figure 2: The dialogue box for defining variable groups.

The DEFAULT group is not modifiable. The usage of groups is described in Section 1.6.

1.5 Adding and deleting random variables (chance nodes)

New random variable represented by a node can be created using the right palette (it can be opened/closed by clicking the arrow icon in the top-right corner of the main window). The class of the necessary item can be selected in the palette, the cursor will change its shape, and new instances can be created by clicking in the main window. By pressing the ESCAPE key, the status of the editor (and the shape of the cursor) will change back to its normal mode. The palette contains chance, decision and utility nodes, with subversions for special local models, and connectors/edges (discussed later respectively).

A node can be deleted by selecting it and pressing the DELETE key.

1.6 Modifying the properties of a variable (chance node)

The PROPERTIES view below the main window shows and allows the modification of the properties of the nodes. By clicking on a node, its properties will be enlisted in this view:

Annotations. For each variable, keyword-free text pairs can be defined in the dialogue box opening by clicking on the textsc... button in the ANNOTATIONS row.

Group. A variable is always assigned to a single group. It is the DEFAULT group after creation, which can be changed to any existing group.

Label. The name of the variable (and node). It can be also modified by selecting the node in the main window and clicking again on its name.

Type. Chance and action nodes always have a type, which can be selected to any existing type.

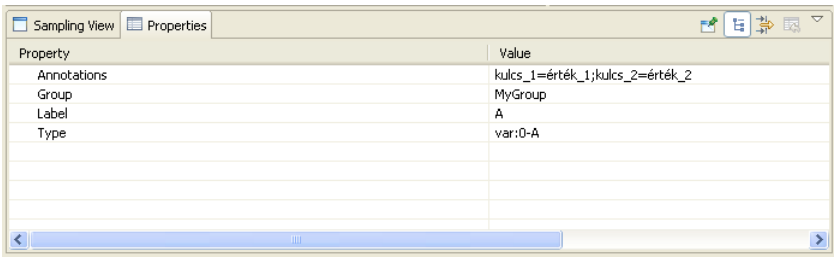

Figure 3: The properties view.

1.7 Adding and deleting edges

Similarly to nodes, edges can be added to the BN structure from the right palette of the main window. By selecting its class, the source and destination can be selected. The structure of a Bayesian network should be a directed acyclic graph, which is monitored and the system does not allow to insert an edge creating a directed cycle.

1.7.1 Local probabilistic models by conditional probability tables (CPTs)

Besides the structural level defining the conditional independences among the variable, there is a quantitative level defining the stochastic dependencies between the the variables. This is routinely specified through local conditional models representing the stochastic dependency of the children from their parents. Assuming that the random variables corresponding to the parents and the child are discrete variables, the most general local conditional model is the "table" model, which specifies a separate multinomial distribution for each parental value configuration. Specifically, if the parents are denoted with $X_1, \ldots, X_k$ with arity $r_1 = |X_1|, \ldots, r_k = |X_k|$ and Y denotes the child with arity $r_Y = |Y|$, then the conditional probability table representing the conditional distribution $P(Y|X_1, \ldots, X_k)$ contains $r_1 * \ldots * r_k * (r_Y - 1)$ free parameters ($|X|$ denotes the number values of variable X). By clicking on a node, its CPT is visualized by showing the $r_1 * \ldots * r_k$ parental configurations in separate rows and the rightmost $|Y|$ columns shows the values of the child Y.

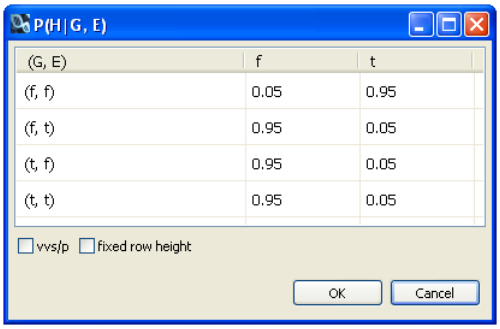

Figure 4: Valószínűségi csomópont feltételes valószínűségi táblájának megjelenítése.

Because each row corresponds to a separate distribution, the sum of the cells should sum to 1, which is monitored and maintained by the system.

1.7.2 Local probabilistic models by conditional probability decision trees (CPDTs)

The conditional probability table model specifies a separate distribution for each parental configuration. However, the individual treatment of the parental configurations is frequently not practical, because the same distribution can be applied for multiple parental configuration (i.e., $P(Y|X_{1:k}1 = x_{1:k}) = P(Y|X_{1:k}1 = x'_{1:k})$ for $x'_{1:k} \neq x'_{1:k}$)). Note that despite such identities, the parents can be still relevant, because the definition of conditional independence requires irrelevance for all values. Decision trees offer a more economic representation, in which an internal node is a univarite test and branching is labeled by the values of the variable corresponding to the internal node, and a leaf contains a conditional distribution $P(Y|X'_1, \ldots, X'_{k'})$, where the variables are the internal nodes in the path to this leaf from the root. Because this tree usually is not a complete tree and for many leafs $k' < k$, the exponential number of parameters in the table model can be decreased linearly (it could be even further decreased to constant using default tables and decision graphs).

The editor of conditional probability decision trees can be opened by clicking on such a node and a CPDT can be constructed using the following operations.

Adding a new internal nodel. A variable is selected from the list of parents shown in the right palette of the CPDT editor main window, then the node is clicked to indicate its intended position (the new node is inserted "above" the clicked node as internal node). The system indicates the applicability of this operation and does not allow multiple use of a node in the same path (the shape of the cursor indicates the possibility of an insertion).

Deleting a subtree. An internal node in the CPDT can be selected and by pressing the DELETE key the whole subtree will be deleted and replaced to a single leaf node with uniform conditional distribution.

Repositioning a subtree. A subtree can be repositioned by simply dragging to its root node to the target node.

Modification of conditional distributions. Clicking on a leaf will open a probability distribution editor dialogue box, which allows the specification of a conditional distribution $P(Y|X'_1, \ldots, X'_{k'})$ (the variables $X'_1, \ldots, X'_{k'}$ are the internal nodes in the path from the root to this leaf).

The CPDTree can be graphically rearranged by draggin the nodes or selected subtrees. Furthermore, by selecting the TREE LAYOUT item for the pop-up menu, the software optimize the layout of the CPDTree.

After inserting a new internal node above a leaf, the new leafs will inherit the conditional distribution of the original leaf.

1.7.3 Action nodes

Action nodes represent actions/interventions, thus they cannot have parents.

1.7.4 Utility nodes

Utility nodes represent functional dependencies (i.e. utility functions), thus only a single value should be specified for each parental configuration (and its range is not confined to $[0, 1]$).

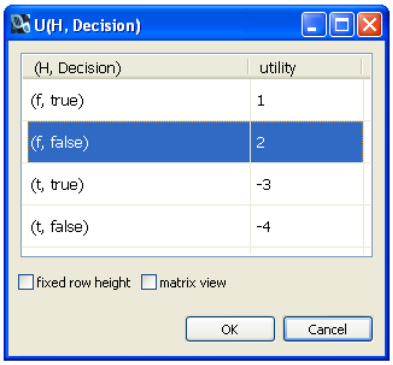

Figure 5: The table representation of a utility function.

In case of two parents, the utility function can be represented in a matrix form, where the horizontal and vertical axis are labeled by values of the first and second parent respectively.

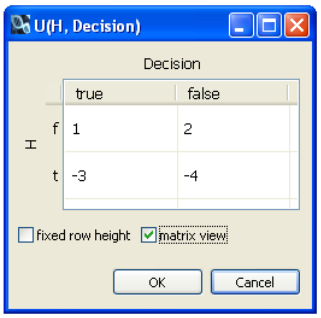

Figure 6: The matrix representation of a utility function.

1.7.5 Decision tree based utility nodes

Analogously to chance nodes, the local dependency model of a utility node can be represented more efficiently using the decision tree representation. The same editor can be accessed with the only difference that the leafs contains a single value (and not a distribution as for chance nodes).

1.8 Format of data files containing observations and interventions

A data file containing observation and interventions matched to the chance and action nodes in the BN. The data file should be a comma separated values (csv) with the following properties.

- The in-line separator is a comma.
- The first line is the header, contianing the names of the variables (optionally with additional variables).
- Subsequent lines should contain values within the intervals of the values of the corresponding variables or an empty cell.

1.9 Setting the BN parameteres from a data set

The parameteres of a BN can be automotically derived from a data set by selecting the Operations > Learning parameters ... menu item and selecting the data file.

2 Visualization of structural aspects of exact inference

The system contains the (*propagation of probabilities in trees of cliques*) algorithm, which is a popular exact inference method for discrete BNs [**?**]. The PPTC method exploits the existence of linear time-complexity inference methods for tree structured BNs, thus it constructs a tree with merged nodes (mega-nodes) from the original BN as follows.

Moral graph. Create cliques for each parental set, then drop the orientation of the edges.

Chordal graph. The moral graph should be triangulated, i.e. any chordless cycle has at most three nodes (also known as triangulated or decomposable graphs, subsets of perfect graphs).

Clique tree. Merging the maximal cliques to mega-nodes, a special clique tree is contructed.

The efficiency of the inference depends from the properties of this tree, e.g. the number of values of the mega-nodes. The system allows the tracking of the effect of a structural aspect of the original network through this steps, to understand its final effect on the efficiency of the inference.

2.1 Visualization of the edges (BN)

The menu item Operations > Show connections switches on/off the visibility of the original Bayesian network.

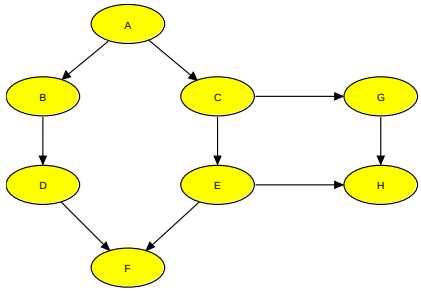

Figure 7: Visualization of the directed edges in the original Bayesian network.

2.2 Visualization of the chordal graph

The menu item Operations > Show chordal graph switches on/off the visibility of the undirected edges of the chordal graph.

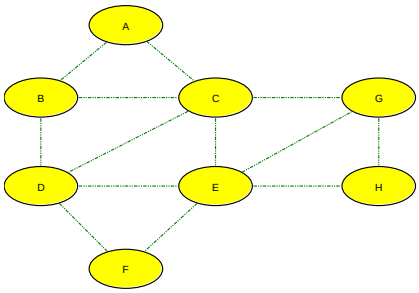

Figure 8: Visualization of the derived chordal graph.

2.3 Visualization of the clique tree

The menu item Operations > Show clique tree switches on/off the visibility of the clique tree, which shows the following items:

Mega-nodes in the clique tree. The mega-nodes in the clique tree are denoted by white, curved squares. The list of variables merged by a mega-node can be inspected by moving the cursor above it (a tooltip will appear and show this list).

Nodes and mega-nodes. The containment of the original nodes in mega-nodes are also indicated by edges.

"Edges" in the clique tree. The mega-nodes are connected through sepsets, which are visualized by dashed lines.

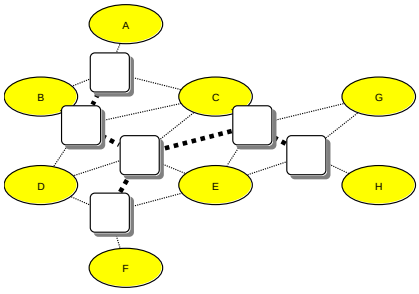

Figure 9: The clique tree of a BN.

3 Inference in Bayesian networks

By selecting the menu item Operations—Inference mode, the *Inference view/Inference specification* opens at the right side of the main window and the *Sampling view/Inference watch* below the main window. These allow the following operations.

3.1 Setting evidences and actions

The GUI supports only "hard evidences", i.e. entering only sure information. By clicking on the bullet of a variable (on the small arrow before its name), a list will open containing its values. By right clicking on the values, the appropriate action can be selected from the pop-up menu, such as Set as evidence, Clear evidence or Clear all evidences.

In case of action nodes only the Set as decision menu item is available to modify the selected value (an action node always should have a selected value).

The selected values are also shown in the node below the name of the variable.

3.2 Univariate distributions conditioned on evidences and actions

After each modification of the evidences or actions, the system always calculates the univariate marginal distributions conditioned on the selected evidences and actions. The conditional distribution of a selected variable is shown at right side of the main window (a variable can be selected by clicking its node or its name in the *Inference specification*). The exact conditional probability can be queried by moving the cursor above the column of the appropriate value.

3.2.1 Inference watch (/Sampling view)

The *Inference watch/Sampling view* allows to track the sequence of inference with varying conditions. In the *Inference specification* view on the right a variable or variable-value pair can be selected to track in the *Inference watch* view. This variable can be the query/target variable, but evidence variables entering into the condition can be also selected. Subsequently, by pressing the Sample button in the *Inference watch/Sampling view* view a new column is created containing the conditional probabilities of the selected values given the current condition, i.e. given the selected evidences and actions. Note that if evidences are watched, then the conditional probalities of their watched values will have probability 0 or 1 (depending that other or this value is selected as hard evidence).

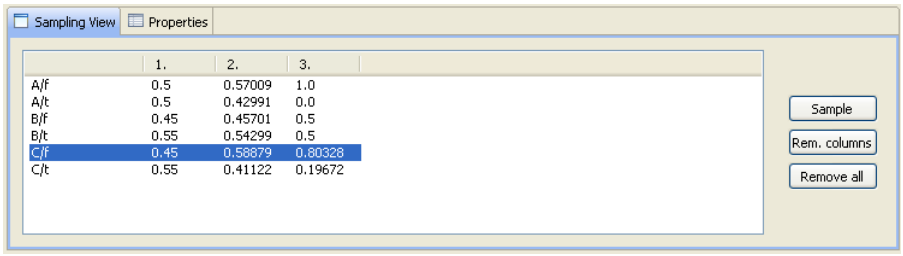

Figure 10: The "Inference watch" allows the tracking of the sequence of inferences with varying conditions.

The *Inference watch* view also supports the following functionalities:

- By clicking any cell, a row can be highlighted and selected, and removed if it is not necessary any more using the REMOVE SELECTED menu item in the pop-up menu (opened by right click).
- The content of the table can be copied to the clipboard by selection the COPY TABLE TO CLIPBOARD AS TEXT in the pop-op menu *(opened by right click), which allows the easy documentation and reporting.
- By pressing the button REM. COLUMNS, the sampled probabilites are deleted, but the selections remains intact.
- The values and the columns can be deleted by pressing the REMOVE ALL button.

3.3 Effect of further information on inference

Further information can significantly change the conditional probabilities, i.e. for a given query and given evidence the conditional probability can drastically if further evidences arise and enter into the condition. The sensitivity of inference to further information provides an overview to this scenario, assuming that the model, both the structure and parameters remains fixed (i.e. not updated by sequentially). In the analyzis a query configuration and initial evidence are fixed, and further information is sequentially entered into the condition, potentially modifying the conditional probabilities of the query. These steps are as follows:

Selection of query configuration. The query value(s) can be selected in the INFERENCE VIEW/SPECIFICATION view by selecting the SET AS SOI TARGET menu item in the pop-up menu (opened by right-click).

Selection of evidences. The evidences and actions are selected in a standard way, as described in Section 3.1.

Selection of further information. Variables as further information can be selected in the INFERENCE VIEW/SPECIFICATION view by selecting the menu item ADD TO SOI CONDITIONS in the pop-up menu (opened by right click). The order of their entry can be modified in the SENSITIVITY OF INFERENCE view using the MOVE UP and MOVE DOWN buttons.

Computation and visualization. By pressing the SHOW button in the SENSITIVITY OF INFERENCE view, the calculations will be performed and an "Information Sensitivity of Inference Diagramm" view will open showin the results.

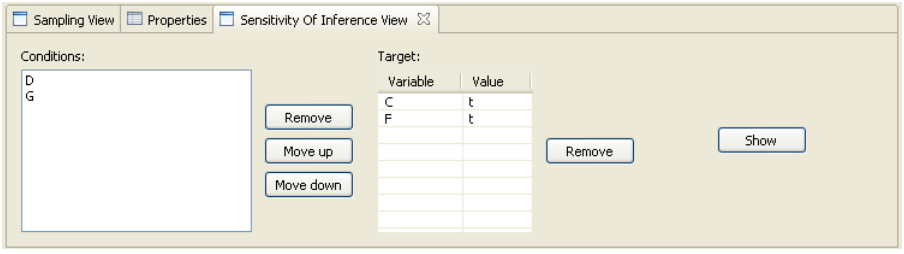

Figure 11: The specification of the analysis of the Information Sensitivity of Inference.

3.3.1 The "Information Sensitivity of Inference Diagram"

In the analysis of the Information Sensitivity of Inference the new informations are set as evidences in the predefined order, i.e. first all the values of the first variable, then all the value pairs of the first and second variables are set as evidence (appropriately always withdrawn the dynamically changing evidences). For each such hypothetical evidence set, the conditional probability of the query/target is always computed. The resulting conditional probabilites can be arranged in a tree, where the root corresponds to $(P(\underline{T} = \underline{t}|\underline{E_0} = \underline{e_0}))$, the children of the root corresponds to $(P(\underline{T} = \underline{t}|\underline{E_0 \cap \{C_0\}} = \underline{e_0 \cap \{c_0\}}))$, where C_0 is the first in the specified order, etc.

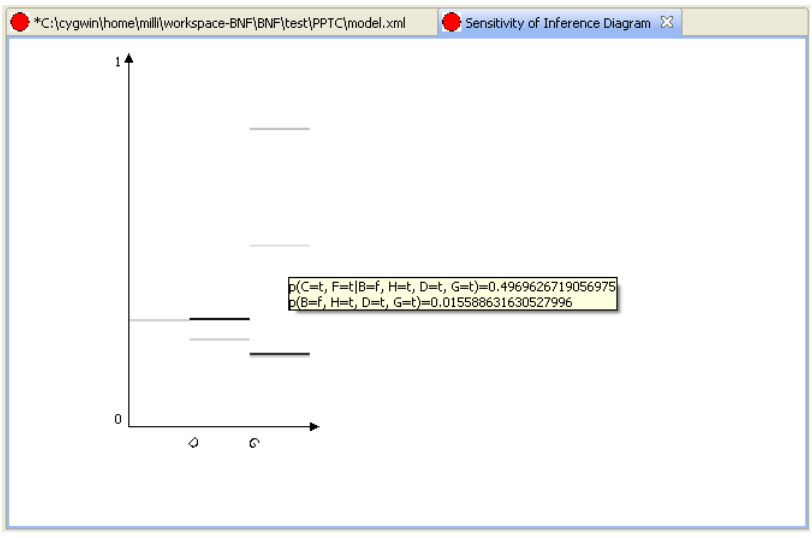

Figure 12: The Information Sensitivity of Inference Diagram.

The Information Sensitivity of Inference Diagram horizontally shows this tree, from its root on the left to the right. The levels of the tree form the columns, and a line in each column represent a pair of probabilities $(P(e_i|e_0), P(teE_i))$: the line is darker for larger $P(e_i|e_0)$, and its vertical position is related to the conditional probability $P(t|e_i)$.

4 Visualization, analysis and modification of the estimated conditional probabilities

The specification of the local probabilistic models in the discrete case with multinomial models can be achieved by the direct estimation of the conditional probabilities. Because of the human estimation biases the estimation of point probabilities is error prone and the system supports the specification of an *a priori* distribution over the parameters using the Dirichlet distribution with "hyperparameteres", where the "hyperparameteres" can be interpreted as previously seen "virtual" sample size. The "virtual" sample sizes can be specified in the probability distribution editor dialogue box by setting the SAMPLE SIZE/PROBABILITY check box. The point probabilities are automatically derived by computing the maximum a posteriori values.

In case of direct estimation of the point probabilities, human experts can be overconfident or underconfident. Overconfidence means a tendency towards deterministic rules, i.e. tendency to use more extreme probabilities. In contrary, underconfidence means a bias towards uncertainty and "centrality", i.e. tendency to use uniform distributions.

4.1 Visulization of conditional probabilities as BN parameteres

As discussed in Section 1.7.1, conditional probabilities are directly present in the CPTs and in CPDTs in case of multinomial models. The conditional probabilities can be visualized in the *Show probabilities* view, which can be started by selecting the menu item OPERATIONS—SHOW PROBABILITIES. The following visualizations are possible by selecting the type of items for the horizontal axis through SELECT CHART:

Variables The conditional probabilities are shown for each variable in a separate column, allowing a general overview about the range and distributions of the parameteres, specifically about possible biases.

Values of a given variable The horizontal axis represents the values of a selected variable, say X, thus the conditional distributions $P(Y|X_{1:k}1)$ can be visualized together by connecting the conditional probabilities corresponding to the same parental configuration $x_{1:k}$ with the same color.

Parental values of a given variable This is the transposed of the "Values of a given variable" case: the horizontal axis represents the parantel value configurations and each column shows the appropriate conditional distribution.

4.2 Biasing the parameters

The effect of the various estimation biases can be emulated in the ESTIMATION BIAS view. It can be started by selecting the menu item OPERATIONS—ESTIMATION BIAS (see Fig. 13).

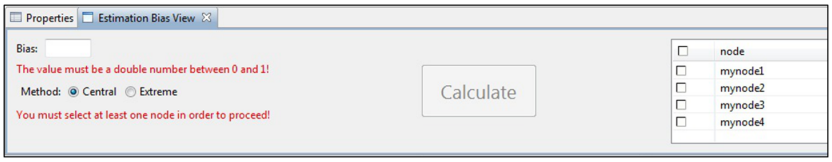

Figure 13: The Estimation bias view.

In the Estimation bias view, the measure/strength of the bias can be defined by specifying a scalar in the interval $[0, 1]$ in the BIAS edit box (1 corresponds to the maximal bias). The type of the bias can be select in the METHOD selection: EXTREME (overconfident) or CENTRAL (underconfident). The list on the right allows the selection of variables, whose conditional probabilities in their local probabilistic models will be modified. By pressing the CALCULATE button, the specified bias is applied.

5 Sample generation

Sample data file can be generated from the BN using the SAMPLE GENERATION view. It can be started by selecting the menu item OPERATIONS—SAMPLE GENERATION. In the dialogue box the output file (path and file name), the type of the output values ("Output type") and the number of complete samples should be specified. For structure learning, please select "Indices" from the pull-down list of "Output type".

6 Structure learning

The system supports the automated construction of a Bayesian network best explaining a given data set (for the restricted case of setting the parameteres based on a data set for a fixed BN structure, see Section 1.9). The system illustrates the learning process using the K2 algorithm, which is based on the idea of the constructive proof for the exactness of the constructed BN: for a given order/permutation of the variables the parental sets are directly defined by their relevance. Without background domain knowledge, the K2 algorithm randomly draws permutations and constructs a BN based on this definition (for the exact scoring and search heuristics, see Cooper, Herskovits(1992)).

The structure learning can be started as follows:

- Select the menu item OPERATIONS—STRUCTURE LEARNING.
- Select data file, set the maximum number of parents (e.g. 3), and the number of permutations (e.g. 10^4), then press START.

After starting the learning, a new model is created based on the data and a STRUCTURE LEARNING view is opened below the main window. During the

learning process, the main window is constantly updated and shows the model found so far best explaining the data set and the STRUCTURE LEARNING view shows the following:

- Progress bar (based on the permutations already investigated).
- The STOP button to halt the process.
- The settings below PARAMETERES .
- The properties of the best BN found in the learning process (below RESULTS)..
- The SCORE graph plots for each permutation the score of the best BN found till this step.

The parameteres of the final model are similarly set based on the selected data set, thus it can be applied in standard inference. The visual appearance of the resulting model can be modified by transferring the properties of an earlier BN using the menu item FILE—IMPORT COLOR, POSITIONS FROM MODEL.